GENIE

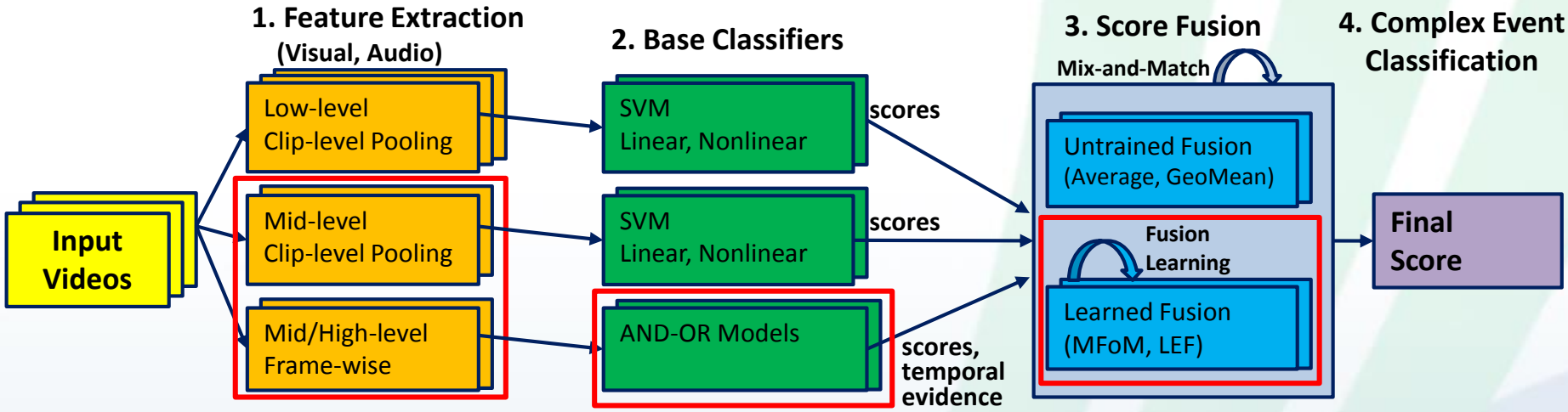
General Engine for Indexing Events

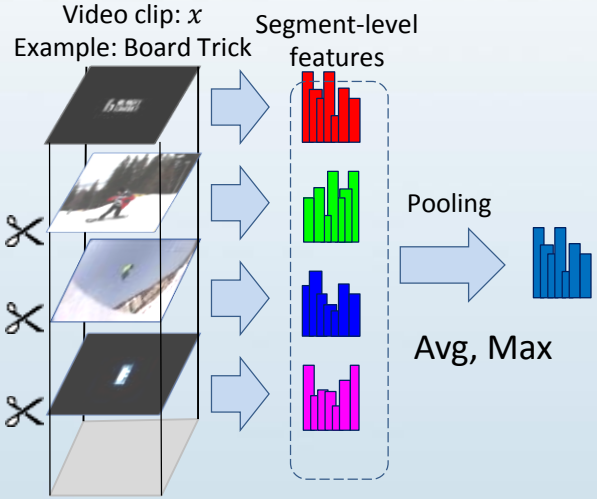
Kitware	Simon Fraser University	Honeywell	Georgia Tech	SUNY Buffalo	Stanford
Sangmin Oh	Kevin Cannons	Ben Miller	You-Chi Cheng	Chenliang Xu	Fei-Fei Li
Ilseo Kim	Hossein Hajimirsadeghi	Scott McCloskey	Zhen Huang	Rohit Kumar	Daphne Koller
Megha Pandey	Arash Vahdat		Chin-Hui Lee	Wei Chen	Vignesh Ramanathan
Amitha Perera	Greg Mori			Jason Corso	Kevin Tang
					Armand Joulin
					Alexandre Alahi

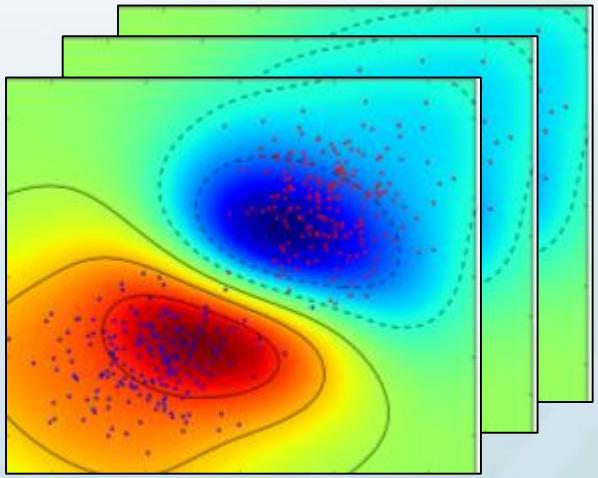
This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 and by the Defence Advanced Research Projects Agency (DARPA) under contract number HR0011-08-C-0135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, DARPA, or the U.S. Government.

Multimedia Event Detection

GENIE 2013 System

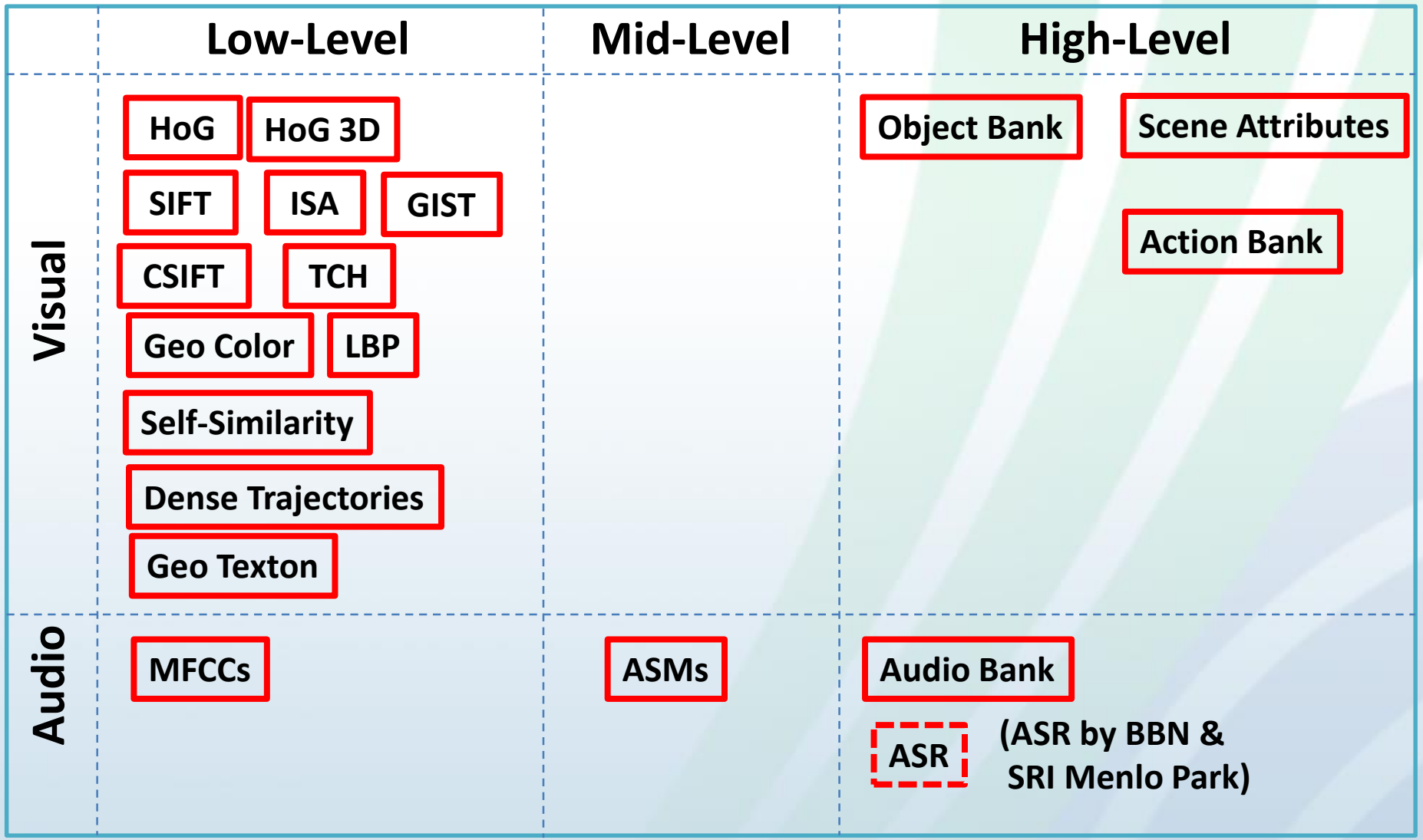






Oh, McCloseky, Kim, Vahdat, Cannons, et al, in MVA Special issue on Multimedia Event Detection, 2013. "Multimedia Event Detection with Multimodal Feature Fusion and Temporal Concept Localization"

2013 GENIE Features



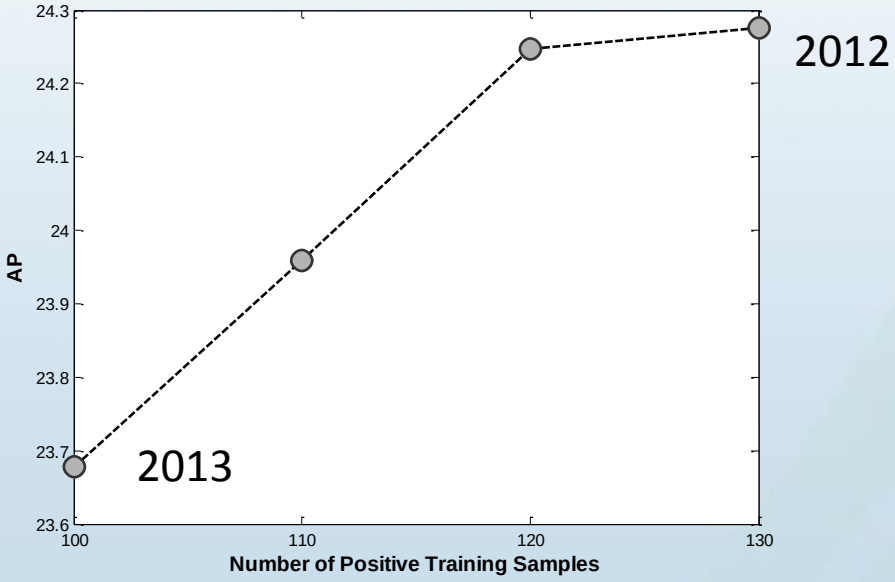
MED Results

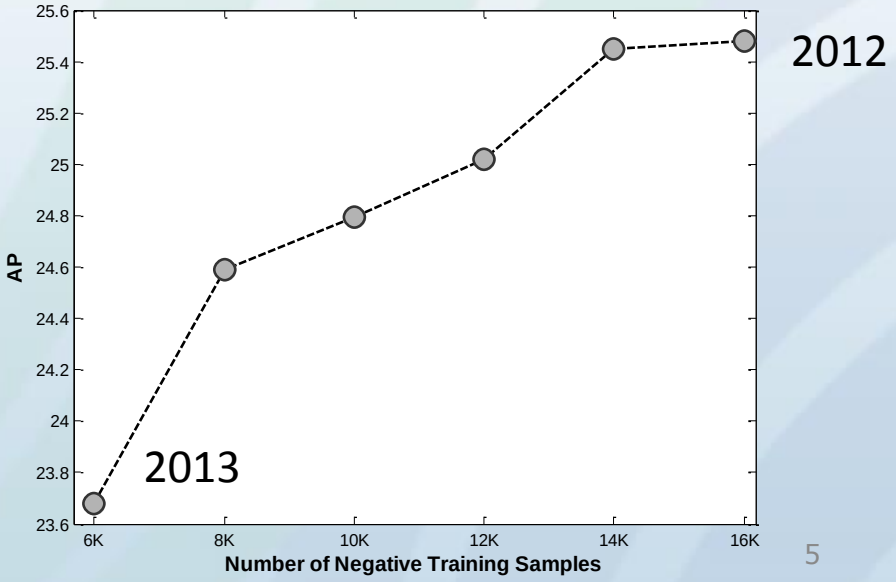
- MED 2013 / 2012

100Ex Full				10Ex Full				0Ex Full PS	
2012 PS	2013 PS	Diff	2013 AH	2012 PS	2013 PS	Diff	2013 AH	2013 PS	2013 AH
23.9	23.3	-0.6	20.2	7.7	10.4	2.7	11.7	1.3	0.4

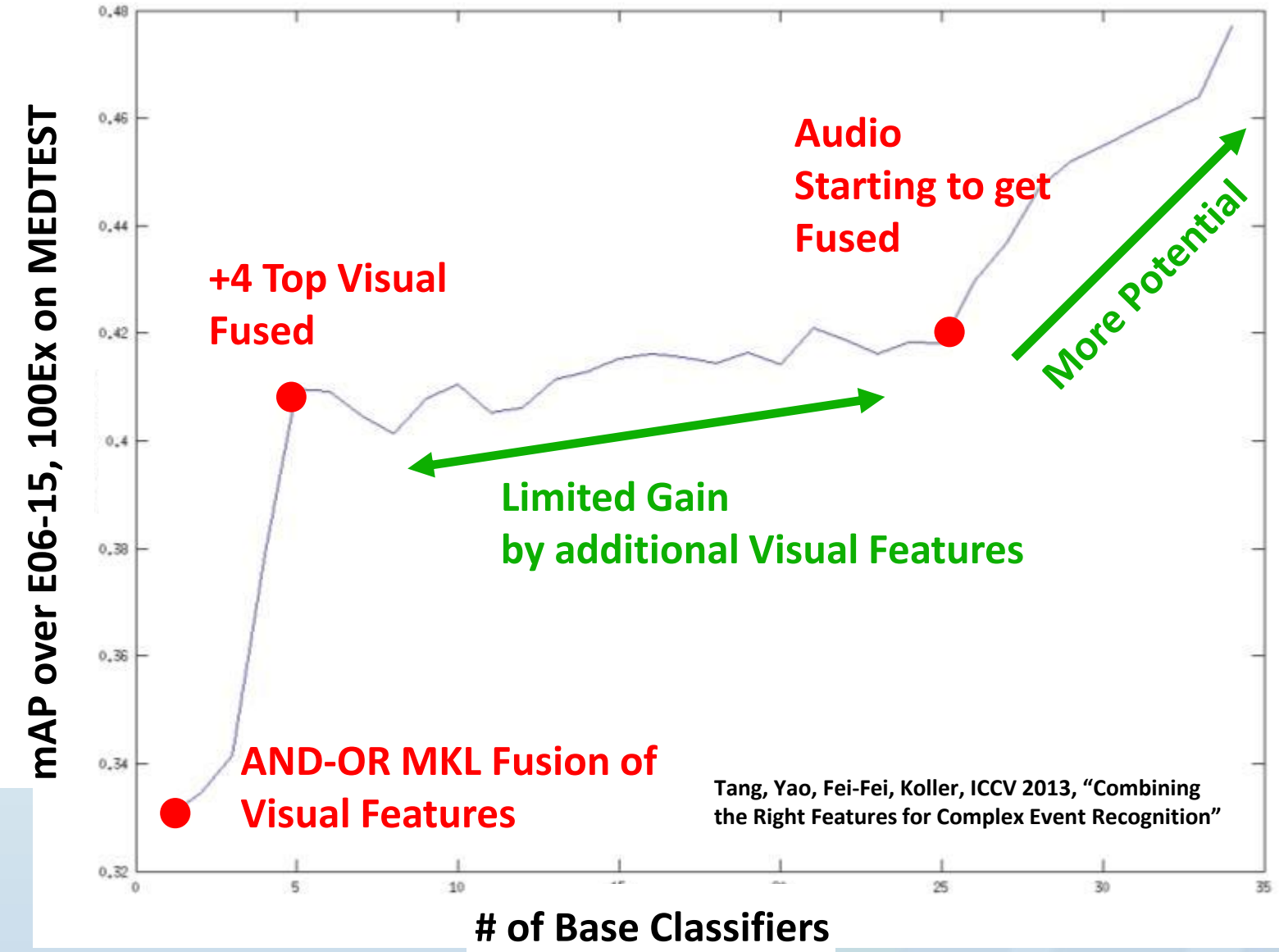
Intrinsic Performance of our system improved.

Performance by a same base classifier with different settings 2012 vs 2013
- Results on MEDTEST '13

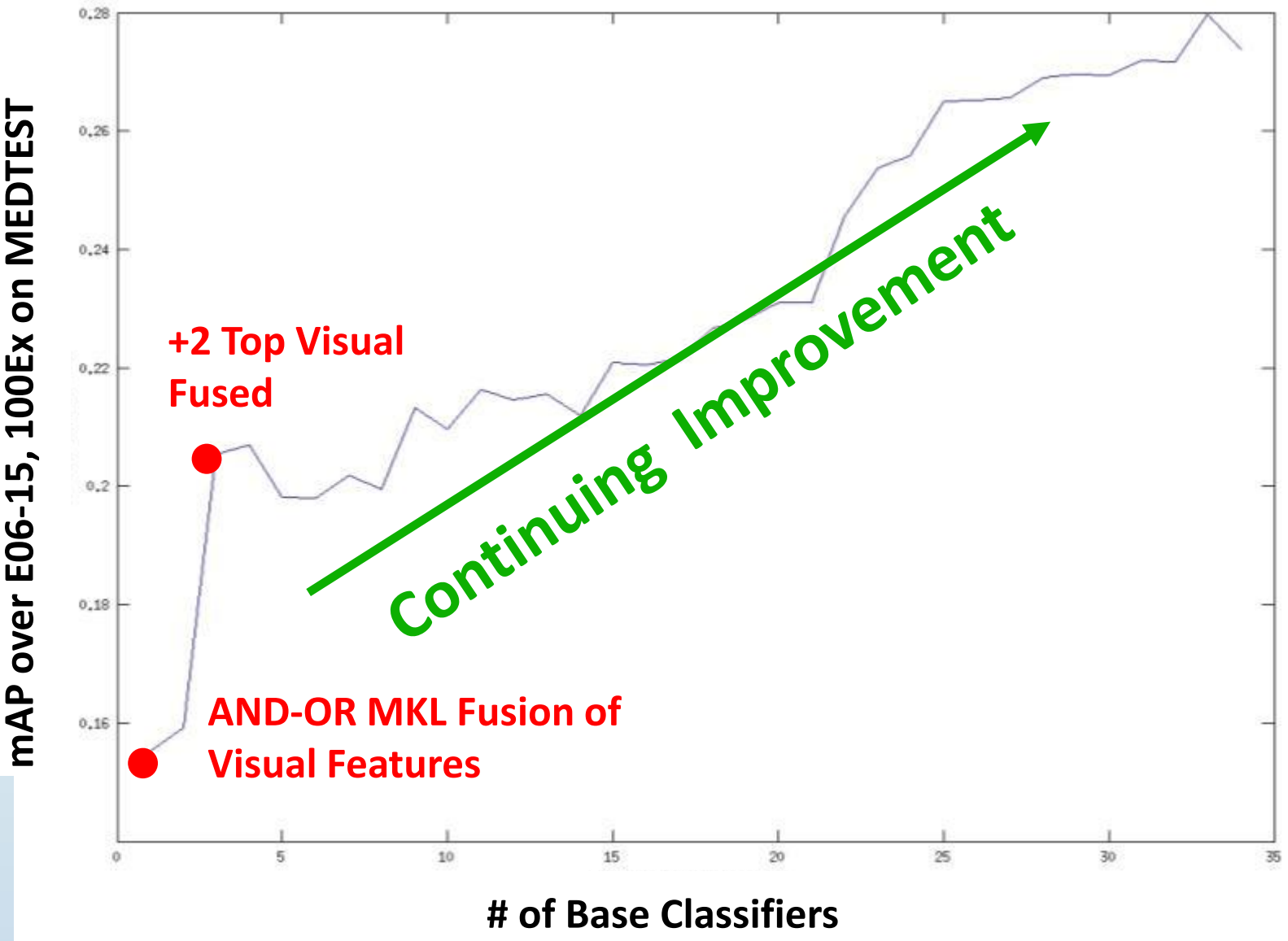




100 Ex Large-Scale Score Fusion



10 Ex Large-Scale Score Fusion



Highlight: 0Ex Audio Bank (56 Concepts)

Human Voice, **Speech**, 1 Person (Eng), Many (Eng), 1 Person (Not Eng), Many (Not Eng), **Other Human Voice**, Cheer, Yell or Scream, Cry, Laugh, Children or Babies, **Sing**, Sing without BG music, Sing without BG music

Other Human Made Sounds, **Hand or Foot**, Footsteps, Clap, **Tool Sounds**, Silverware or Dishware, Knocking on a Surface, Wood Colliding, Metal Colliding, Chopping, Sawing,

Music, Album-like, With Voice, Instrument Only, Soft and Lyric, Heavy/Rock/Excited, Other Music Genre,

Animal, Dog, Cat, Horse, Bird, Beast Roar

Machine Sounds, Light (Appliance, dishwasher), Strong (Electrical Saw, Driller), Motor Vehicle, Aircraft

BG Sounds, Wind, Traffic, Crowd, Water, Weak Background Music, Alarm, Radio, Fry, Fire, Firecracker, Micro BLOW **Noise**

Highlight: 0Ex Audio

- **Flash Mob**
  - yell_or_scream, footsteps, music, strong (electrical saw, driller), alarm
- **Birthday Party**
  - other_human_voice, cheer, yell_or_screem, laugh, sing (casual), clap
- **Grooming an Animal**
  - animal, water

		FullSys	ASRSys	AudioSys	OCRSys	VisualSys
0Ex	BBNVISER	5.2%	1.4%	0.5%	2.8%	3.5%
	CMU	3.7%	1.8%	0.3%	2.1%	2.4%
	Genie	1.3%	1.7%	1.1%		1.0%
	IBM-Columbia	1.6%		0.2%		1.8%
	SRIAURORA	7.0%	3.0%	0.2%	3.7%	6.5%
	Sesame	2.4%	1.7%		2.3%	1.3%
	TNO					
	UMass	5.6%	2.3%		3.3%	5.1%
	VisQMUL	0.2%				

	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
BBNVISER	8.1%	2.5%	0.6%	3.0%	5.0%
CMU	10.1%	3.1%	0.2%	2.8%	5.2%
Genie	0.4%	0.4%	0.5%		1.2%
IBM-Columbia	1.1%		0.2%		1.3%
SRIAURORA	1.4%	3.9%	0.2%	4.3%	0.6%
Sesame	2.8%	2.2%		2.2%	1.3%
TNO					
UMass	1.0%	2.1%		3.9%	0.5%
VisQMUL	+0.2%		0.2%		+0.2%

Audio 0Ex Approach = Audio Bank + ASR

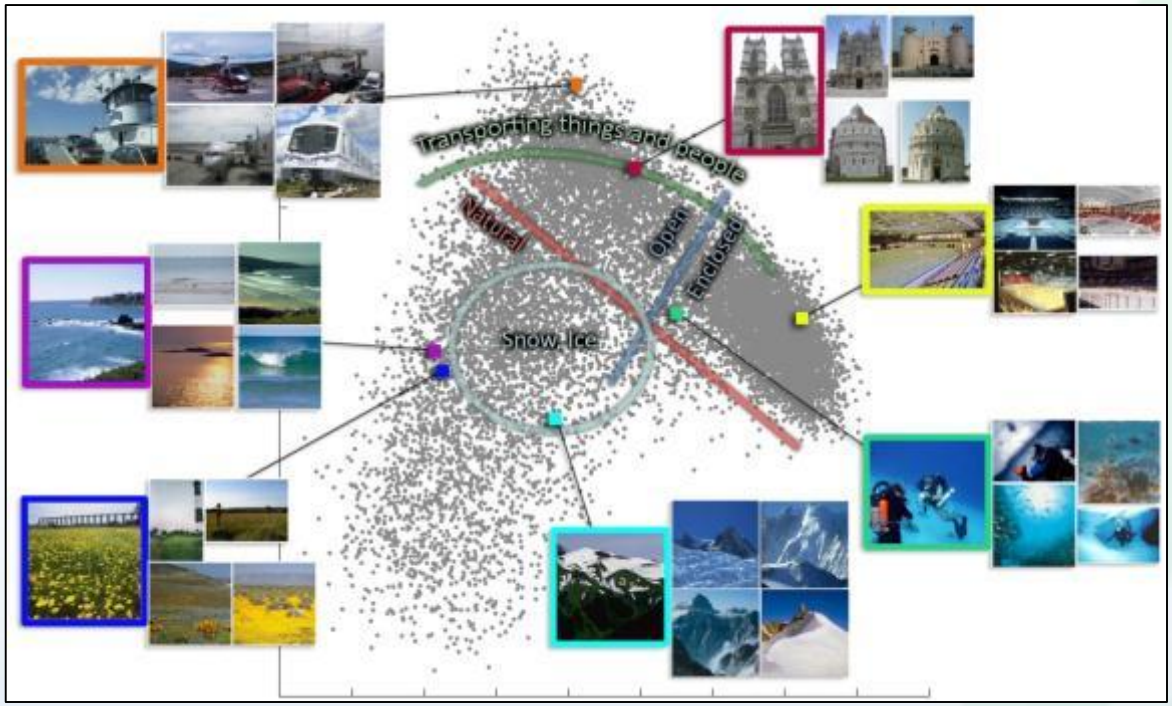
Z. Huang, Y.-C. Cheng, K. Li, V. Hautamaki, and C.-H. Lee. In INTERSPEECH, 2013.
"A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector"

MED Results Online Demo

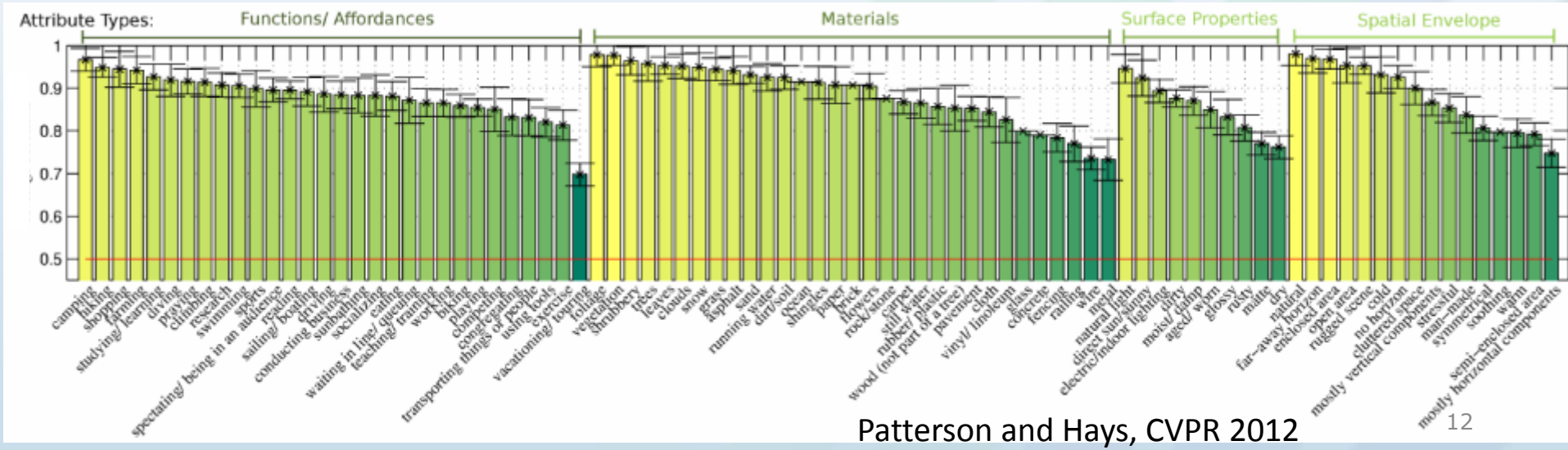
Semantic Concepts

Where we are & Attempts for next step

Scene Attributes



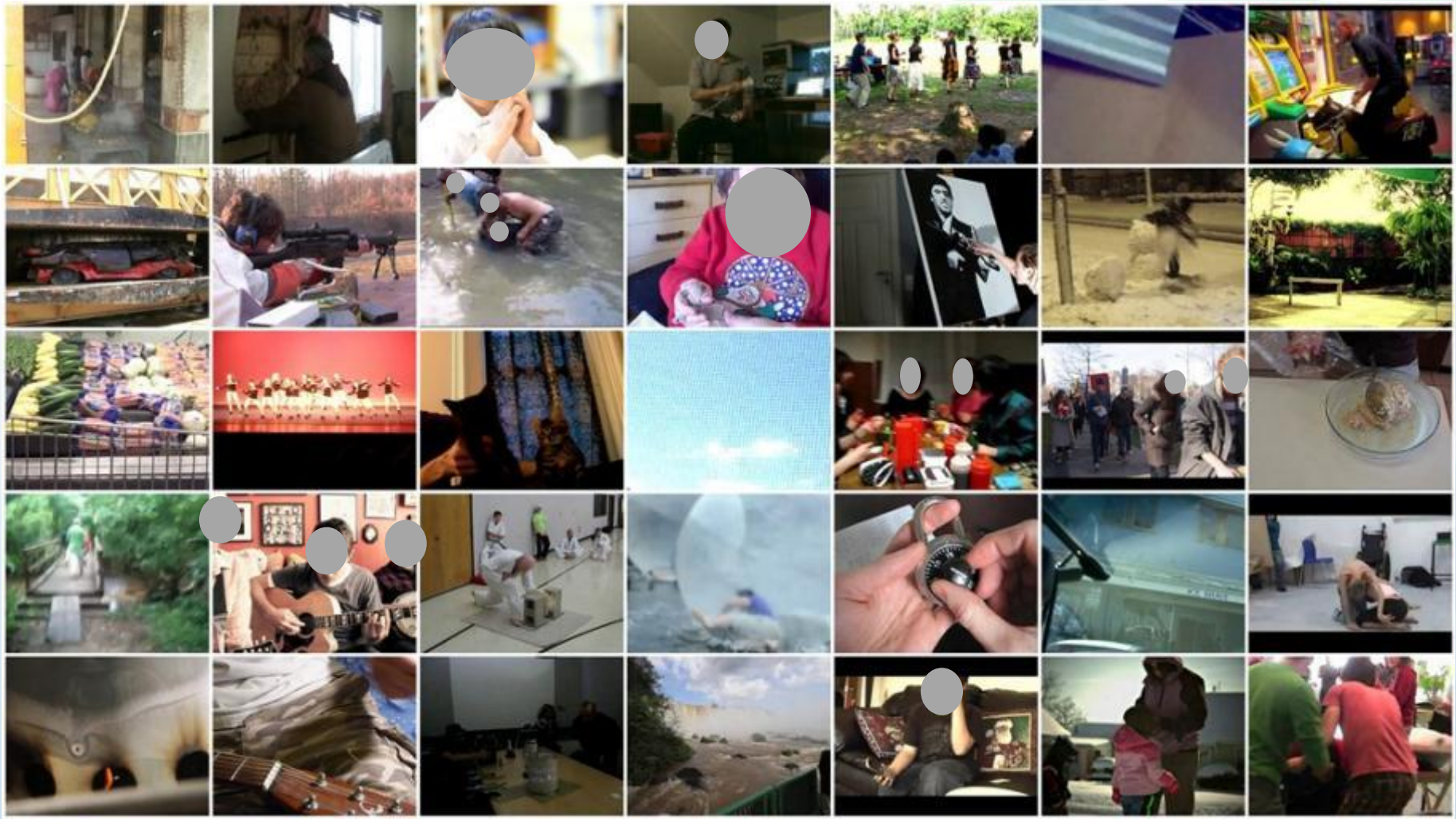
102 Attributes



Patterson and Hays, CVPR 2012

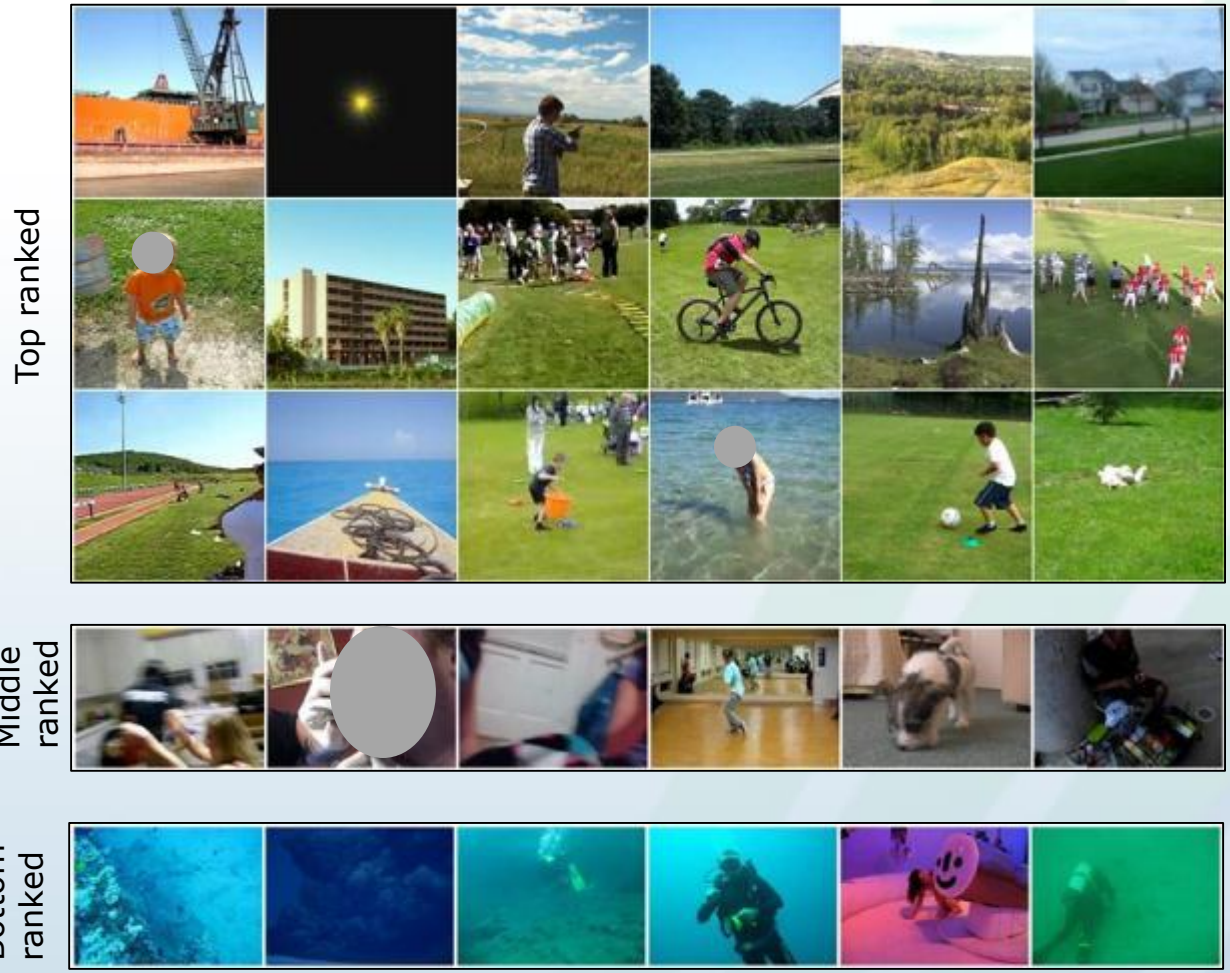
Random Selection

- Randomly selected frames from MED research set (nearly 588, 000 frames extracted from 10,000+ clips.)



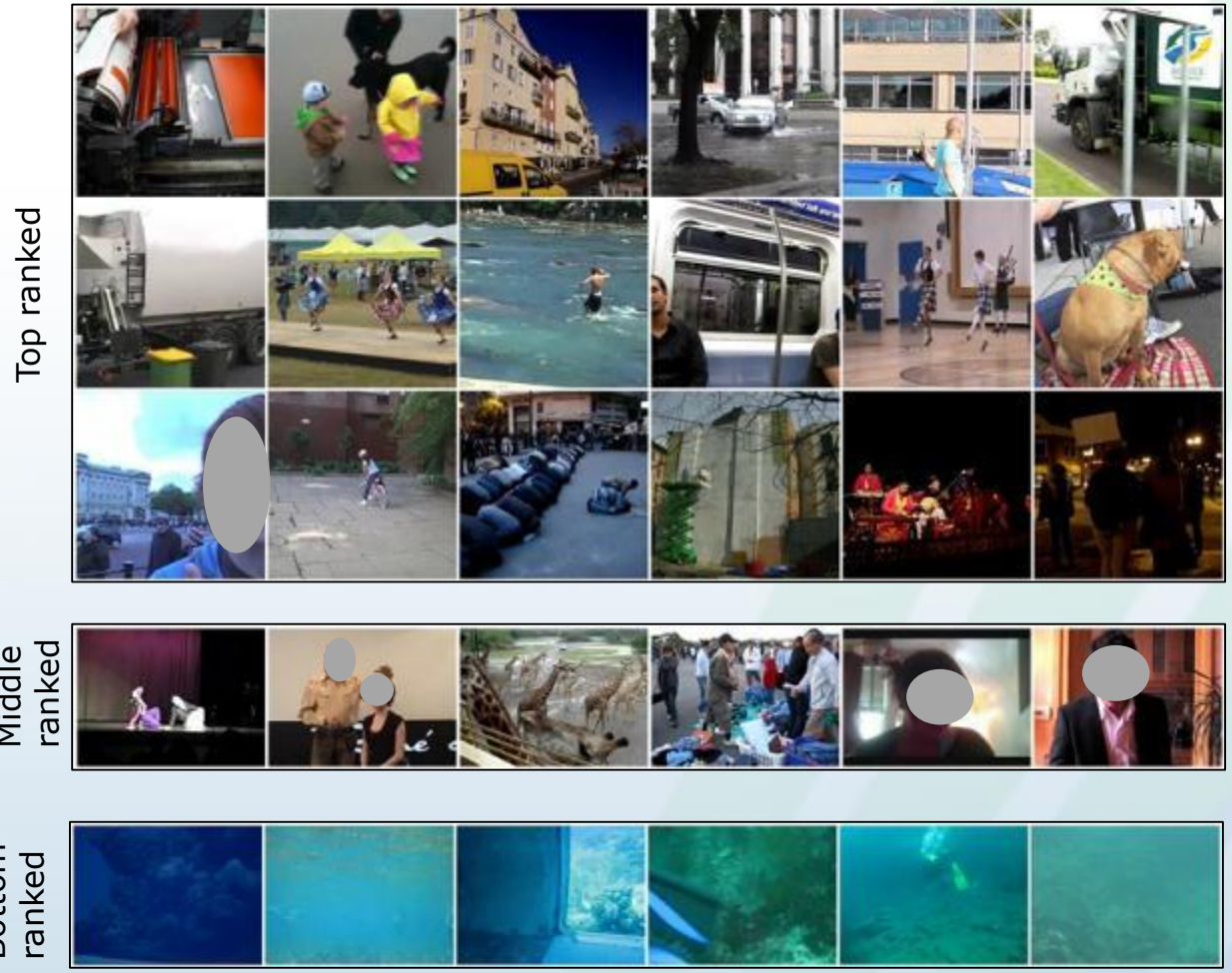
Scene Attributes on MED dataset

- *direct sun or sunny*



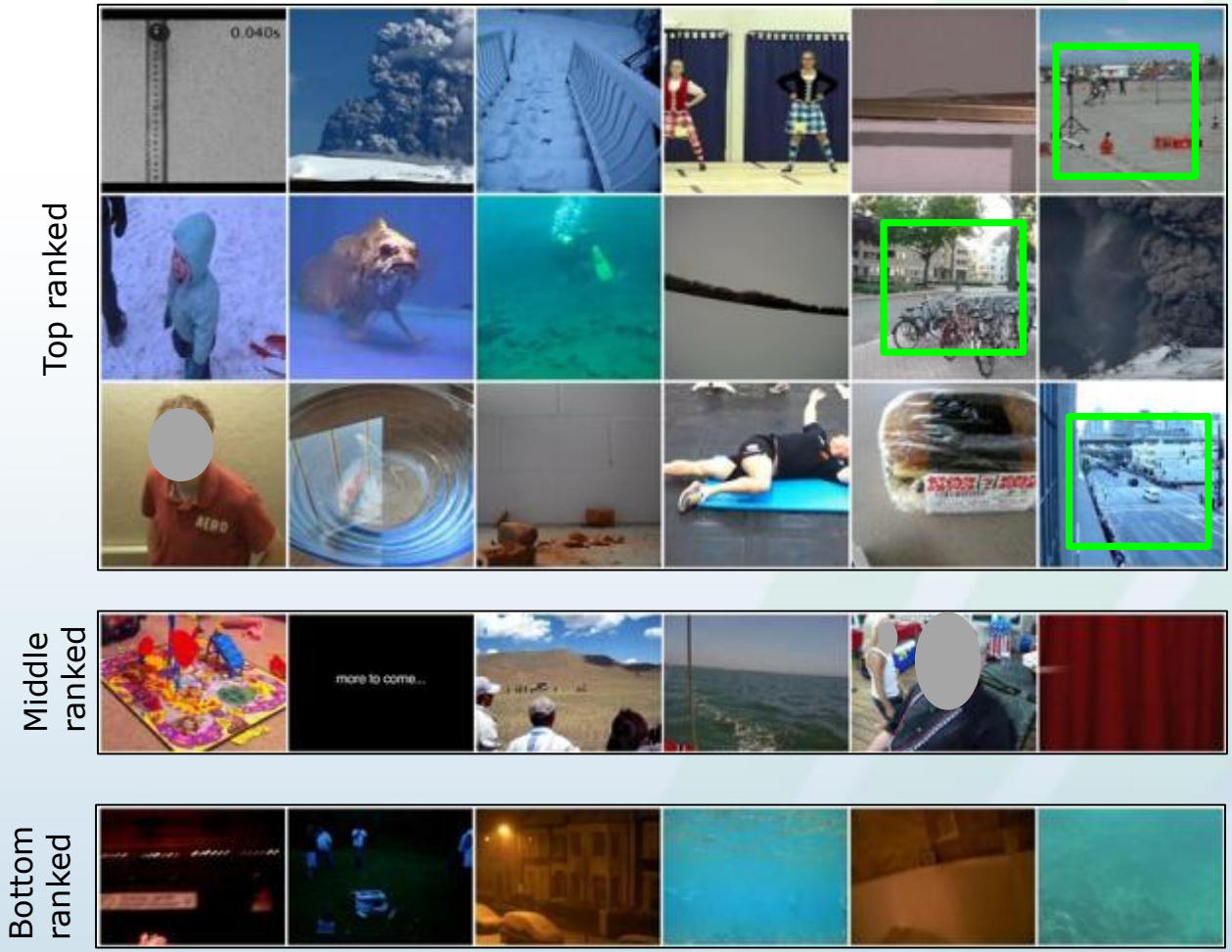
Scene Attributes on MED dataset

- *Man-made*

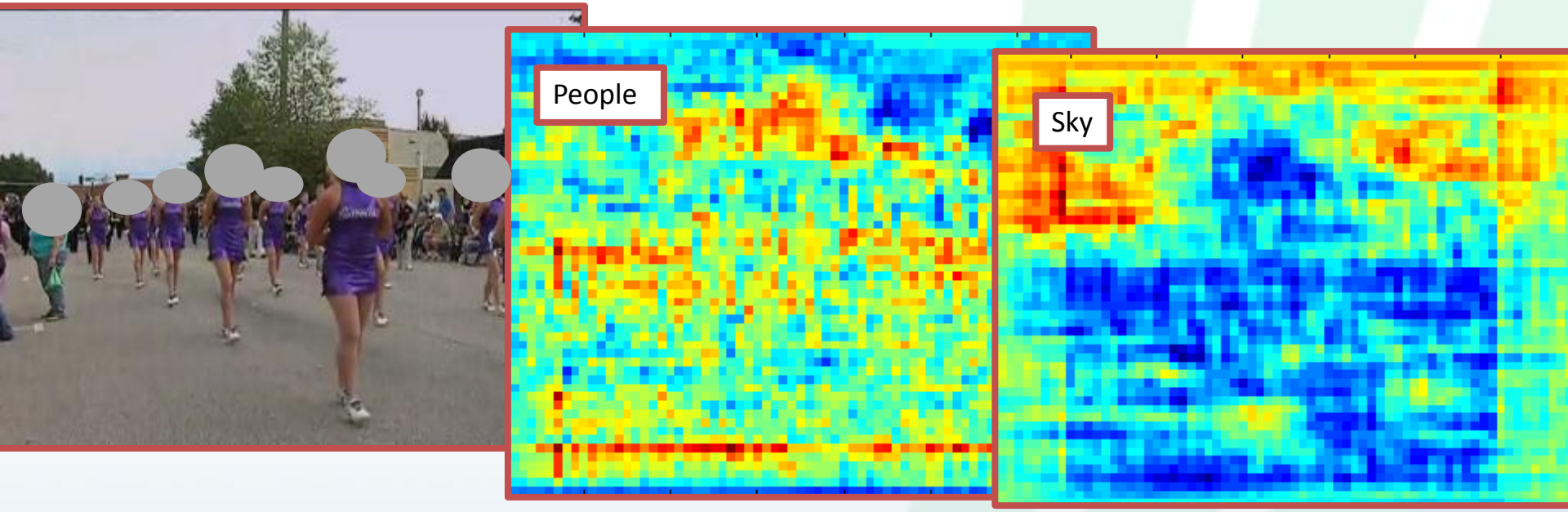


Scene Attributes on MED dataset

- *asphalt* (only a few correct retrievals)



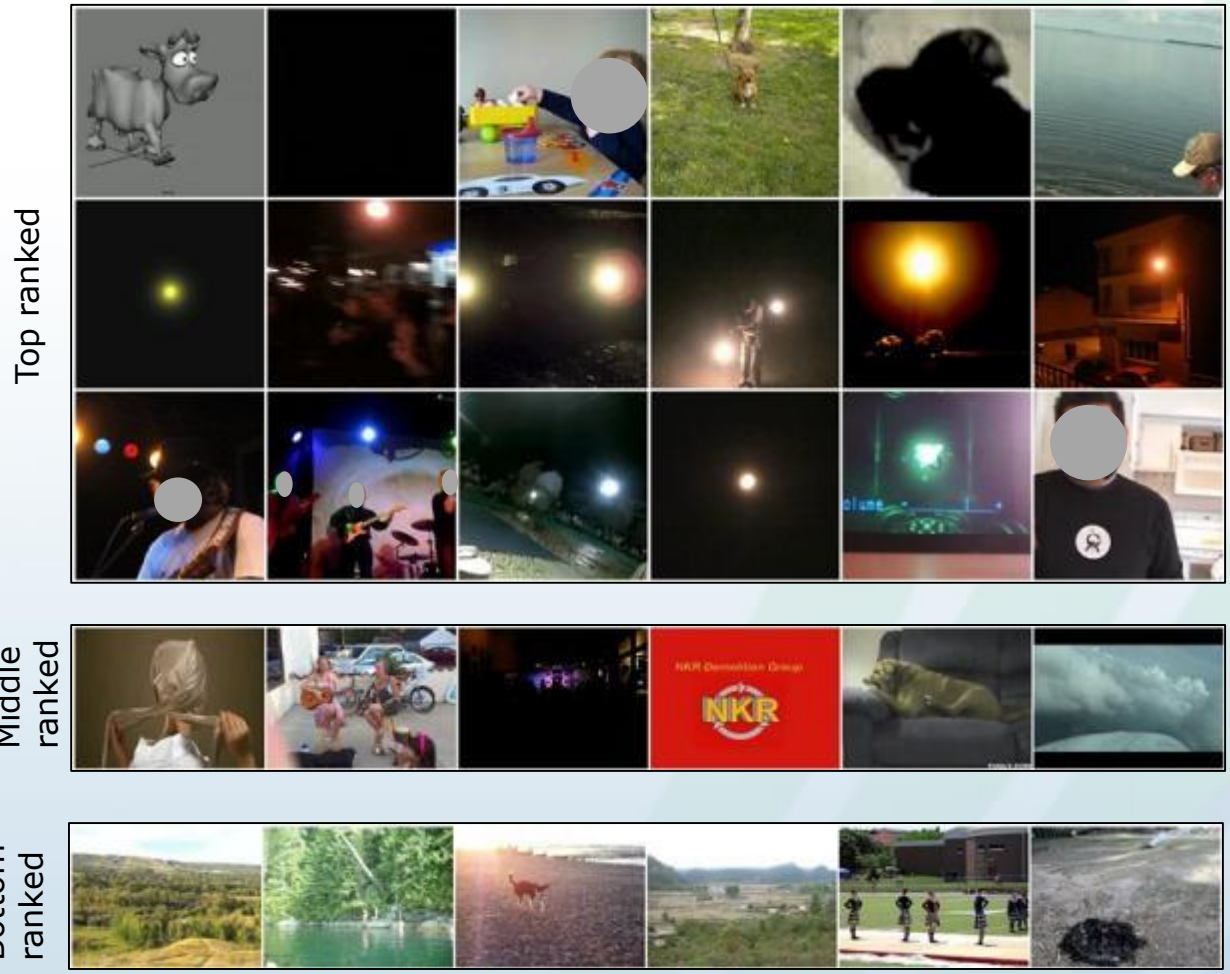
Object Bank



- Li-Jia Li, Hao Su, Eric P. Xing and Li Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification". NIPS, 2010.
- 177 object detectors run at different scales over each frame
  - Computed at key frames
  - 44604-d feature vector, reduced to 177-d by choosing max response per object type
  - Max pooling over all frames
- L2 distance

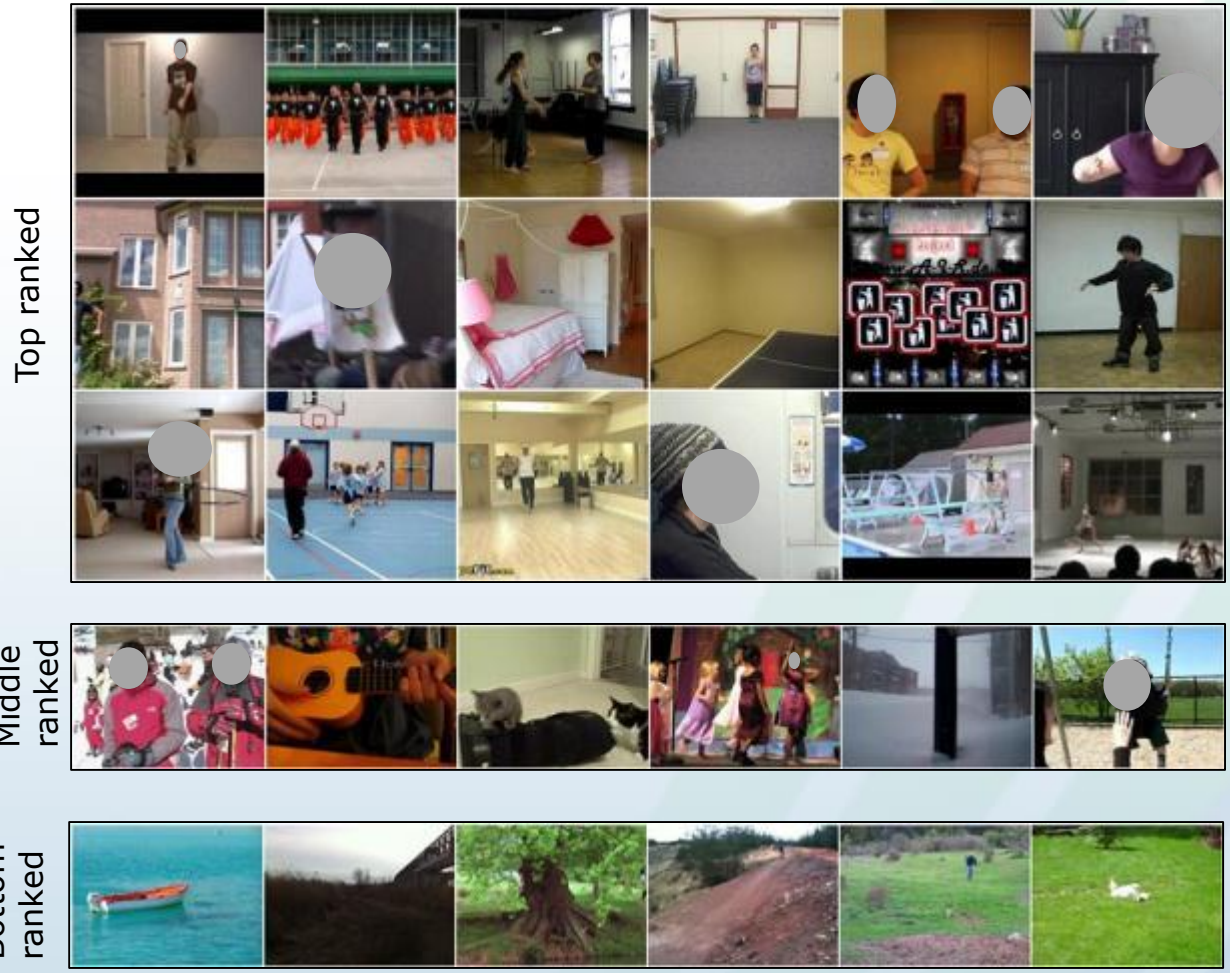
Object Bank on MED Dataset

- *light source*



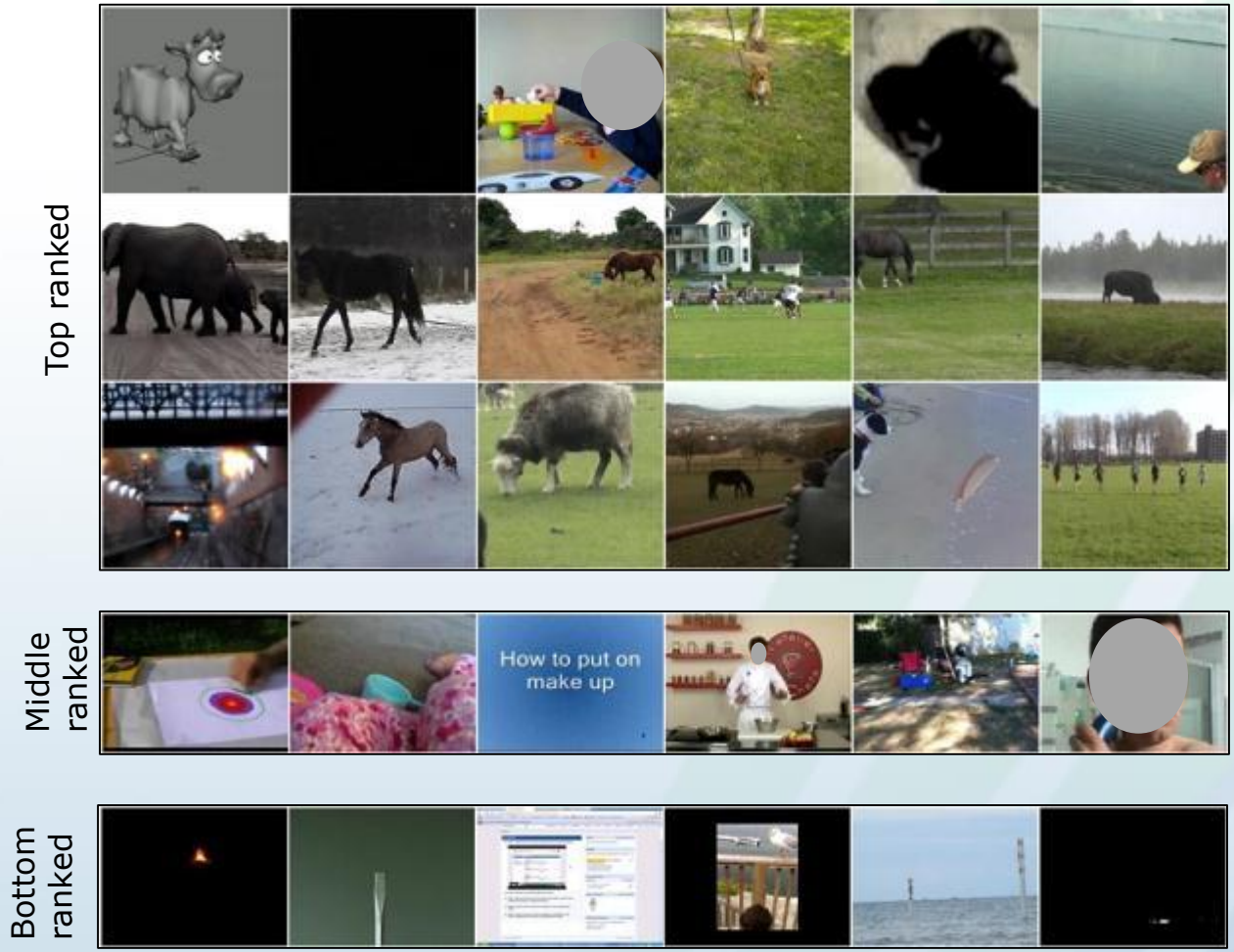
Object Bank on MED Dataset

- *door*



Object Bank on MED Dataset

- *horse* (detects animal or person)



Towards 0Ex Challenges

- How many of visual semantics are reliable?
  - Selected after manual reviews on MED images
    - 107 Scene Attributes → 30
    - 177 Object Bank → 21

- Source of Limitations
  - Imperfect detectors
  - Training dataset is not generalizable for MED
  - Some concepts do not exist in MED

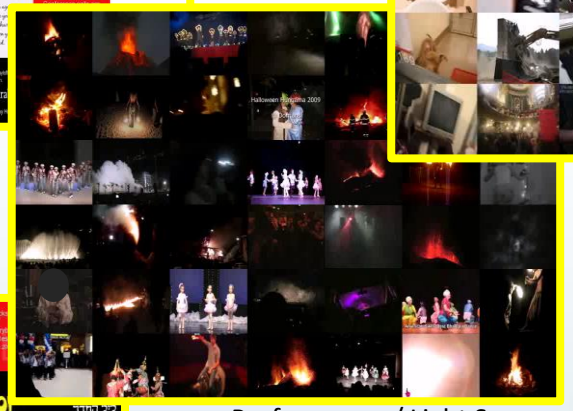
A Solution for Reliable Concepts

- Bottom-Up Concepts
  - Identify concepts that are Easily detectable
    - Simply cluster data
    - Look at each cluster
    - 'Name it' when it makes sense
  - Paradigm shift from 'Top-Down' process
    - Opposite of collecting training data up-front
    - Manual annotation cost incurs at the end
    - Advantage of actually ensuring detectability on your dataset
      - No Transfer learning, and No dataset gap
- Forward Thoughts on Scalability
  - How many concepts can be identified this way?
  - How detailed concepts can be identified?
  - This is an on-going research, but, it looks fairly promising.

Example Bottom-Up Concepts

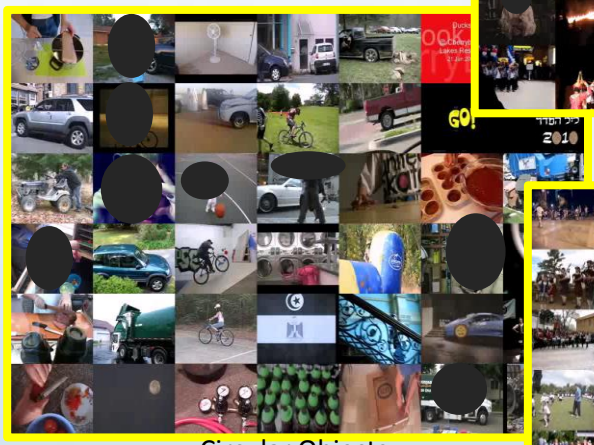
Title / Caption

Hands



Performance / Light Source

Grass / Leaves



Circular Objects

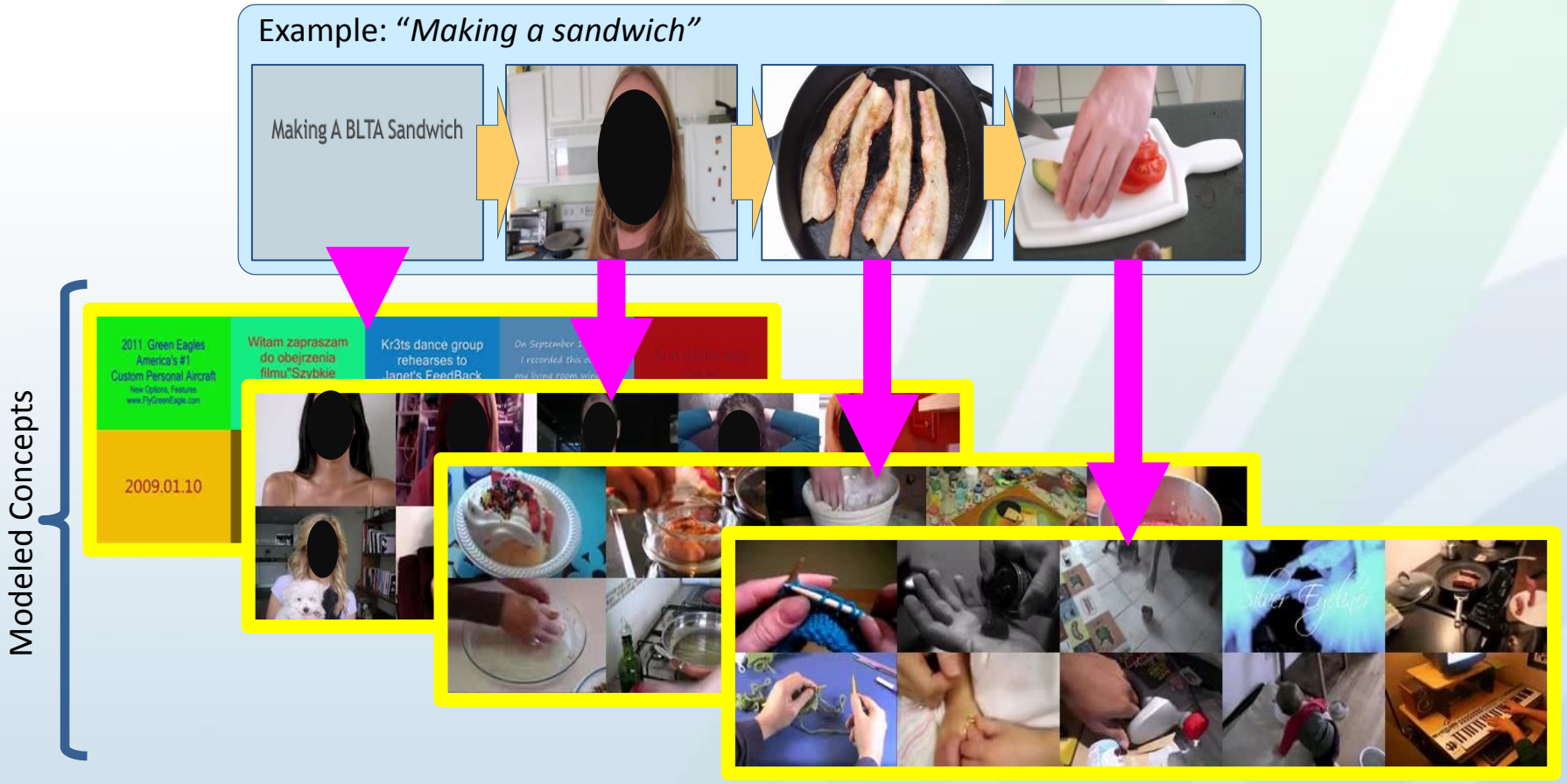
Group of People

Face Close-up

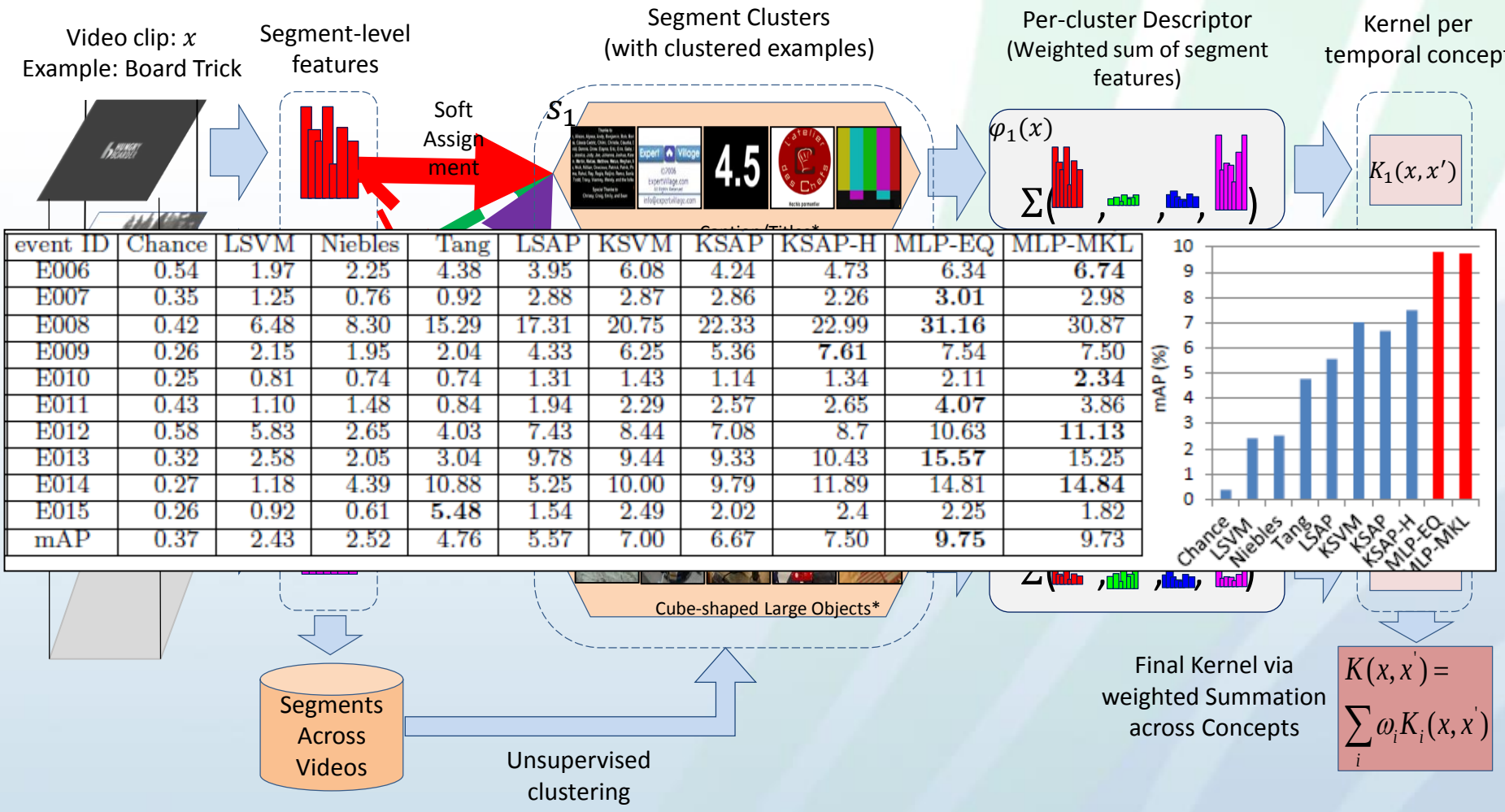
Kim, Oh, Vahdat, Cannons, Perera, Mori,
ACM MM 2013

Capturing Time-Varying Video Contents using Bottom-Up Concepts

- ***We model general semantic concepts*** appearing in real-world videos



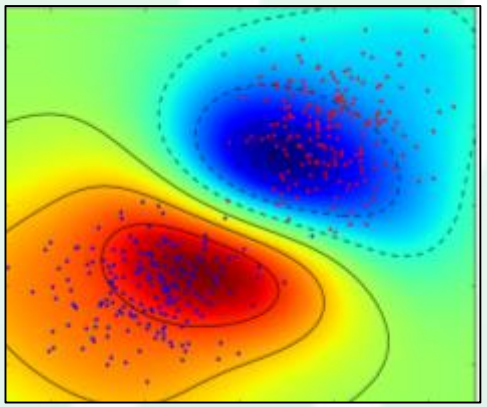
Segmental Multi-way Local Pooling combined with Multiple Kernel Learning



event ID	Chance	LSVM	Niebles	Tang	LSAP	KSVM	KSAP	KSAP-H	MLP-EQ	MLP-MKL
E006	0.54	1.97	2.25	4.38	3.95	6.08	4.24	4.73	6.34	**6.74**
E007	0.35	1.25	0.76	0.92	2.88	2.87	2.86	2.26	**3.01**	2.98
E008	0.42	6.48	8.30	15.29	17.31	20.75	22.33	22.99	**31.16**	30.87
E009	0.26	2.15	1.95	2.04	4.33	6.25	5.36	**7.61**	7.54	7.50
E010	0.25	0.81	0.74	0.74	1.31	1.43	1.14	1.34	2.11	**2.34**
E011	0.43	1.10	1.48	0.84	1.94	2.29	2.57	2.65	**4.07**	3.86
E012	0.58	5.83	2.65	4.03	7.43	8.44	7.08	8.7	10.63	**11.13**
E013	0.32	2.58	2.05	3.04	9.78	9.44	9.33	10.43	**15.57**	15.25
E014	0.27	1.18	4.39	10.88	5.25	10.00	9.79	11.89	14.81	**14.84**
E015	0.26	0.92	0.61	**5.48**	1.54	2.49	2.02	2.4	2.25	1.82
mAP	0.37	2.43	2.52	4.76	5.57	7.00	6.67	7.50	**9.75**	9.73

Segmental Multi-way Local Pooling for Video Recognition,
Kim, Oh, Vahdat, Cannons, Mori, Perera. ACM Multimedia '13.

0Ex Challenges

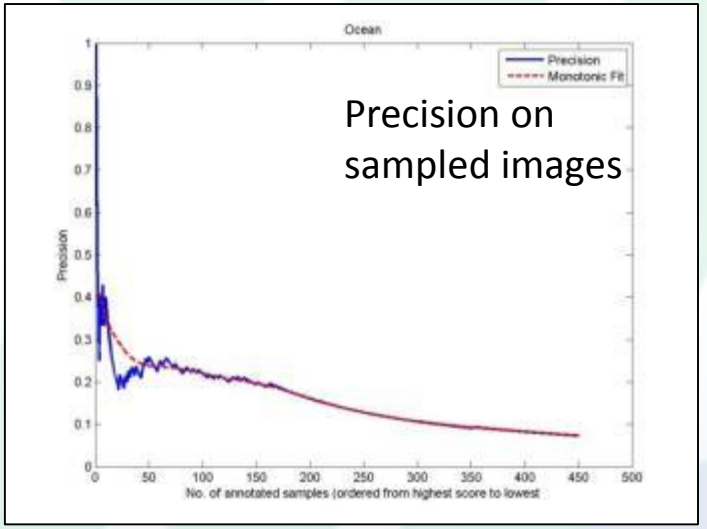


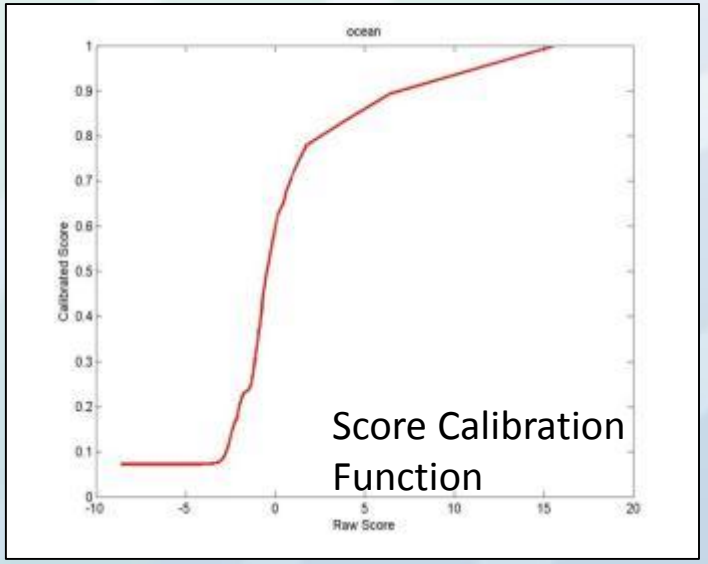
- Detector scores
  - Mostly classifier outputs
    - SVM margins, or transformations
  - Mostly used to rank images/videos
  - Intrinsic meaning of scores are not well-defined
    - Semantic concept classifier scores can be difficult to understand and do not convey true uncertainty to the users.
  - Combining concept scores is an open issue
  - Even, estimating relative strength of multiple concepts on a single input

Semantic Concept Score Calibration

- Goal is to translate the scores into values that are meaningful w.r.t to observed semantics.

- Calibration Process
  - Apply the concept classifiers to the data.
  - Smart sampling of images spanning the complete range of scores. (We sampled ~500 images per-classifier out of ~588, 000 frames extracted from > 10,000 videos.)
  - Manually annotate images w.r.t. relevance of targeted concepts.
  - Estimate precision.
  - Map raw score to estimated precision value.



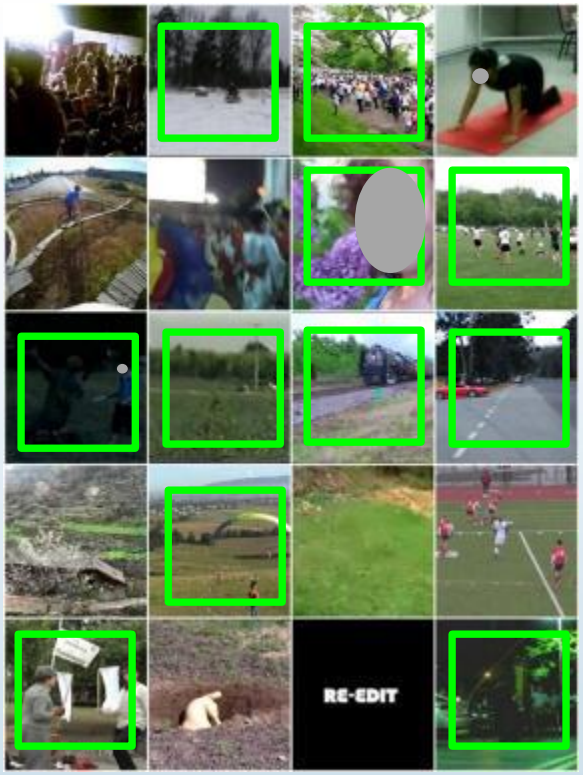


OCEAN

Semantic Concept Score Calibration

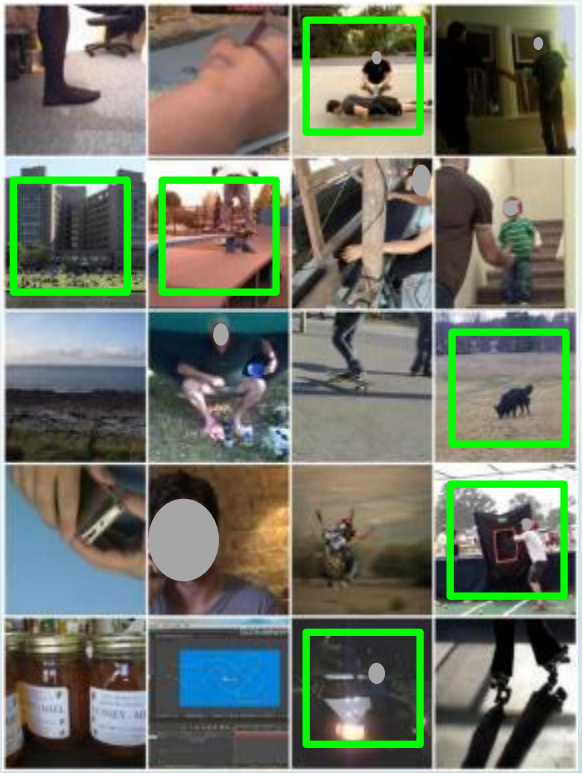
Randomly sample 20 images with *tree* score > threshold (from 100,000+ images)

score > 0.5



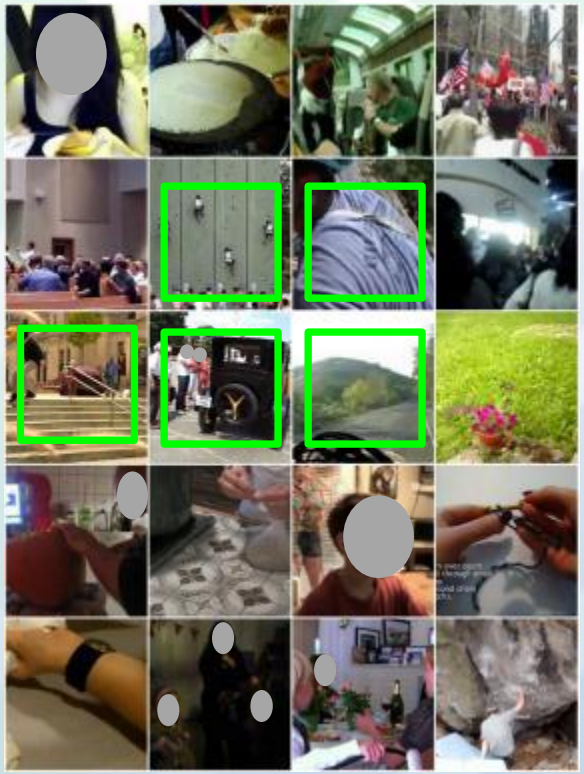
55% correct retrievals

score > 0.3



30% correct retrievals

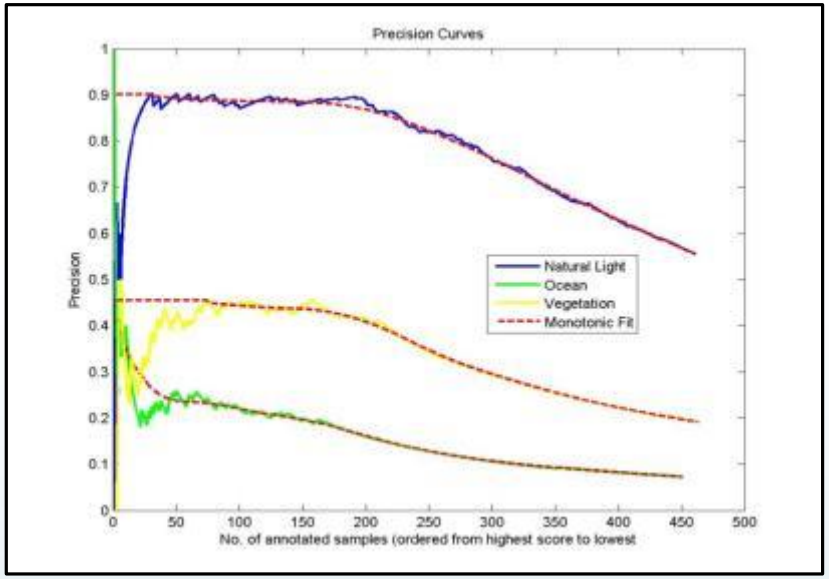
score > 0.2

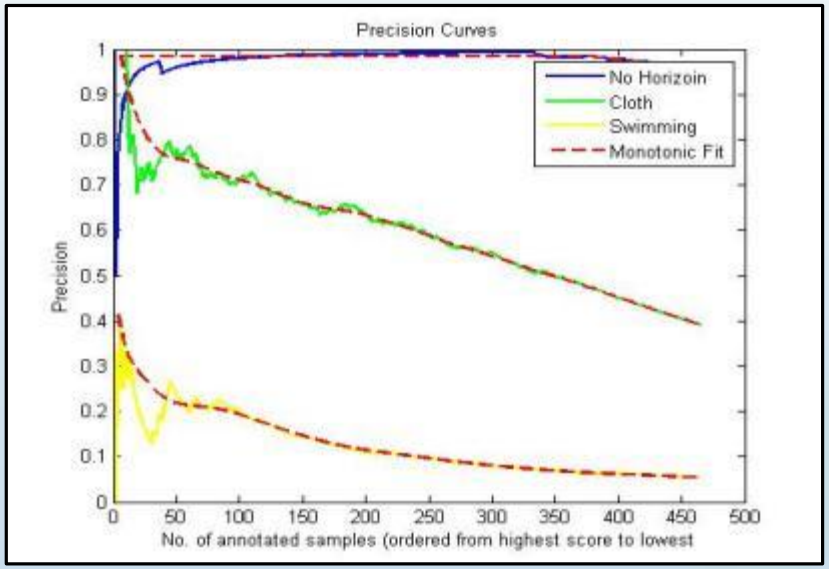


25% correct retrievals

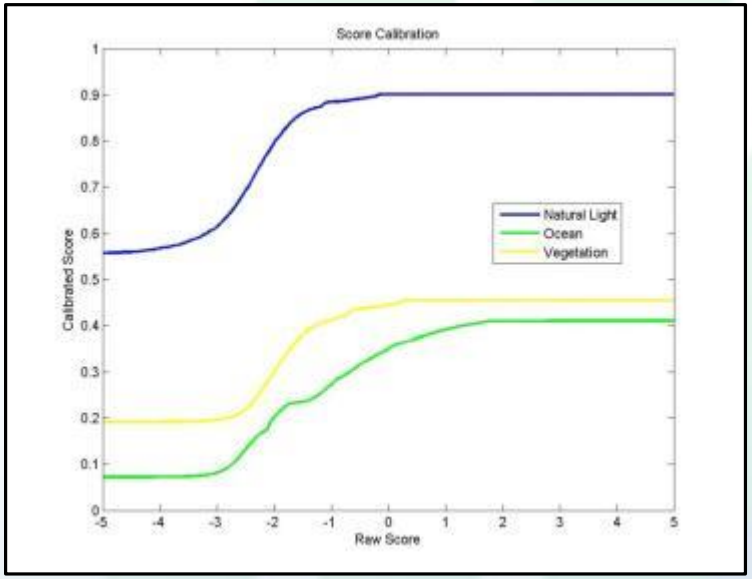
Semantic Concept Score Calibration

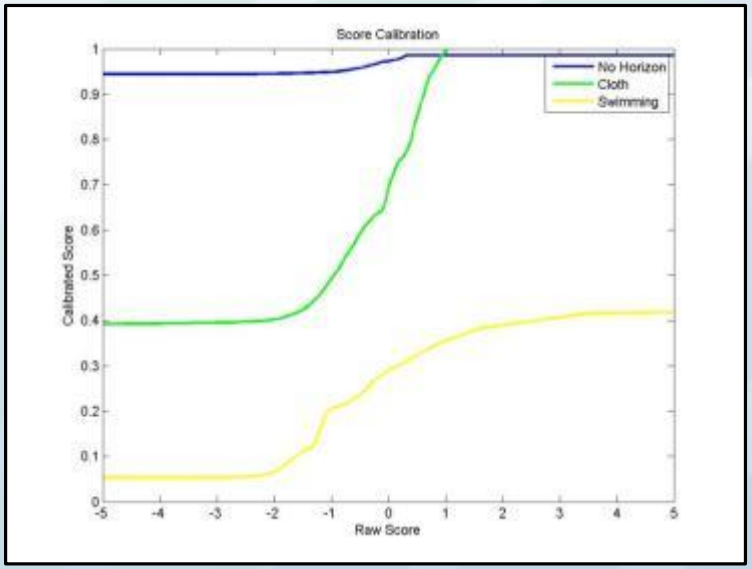
Precision Curves Computed from Annotated Images





Score Transformation



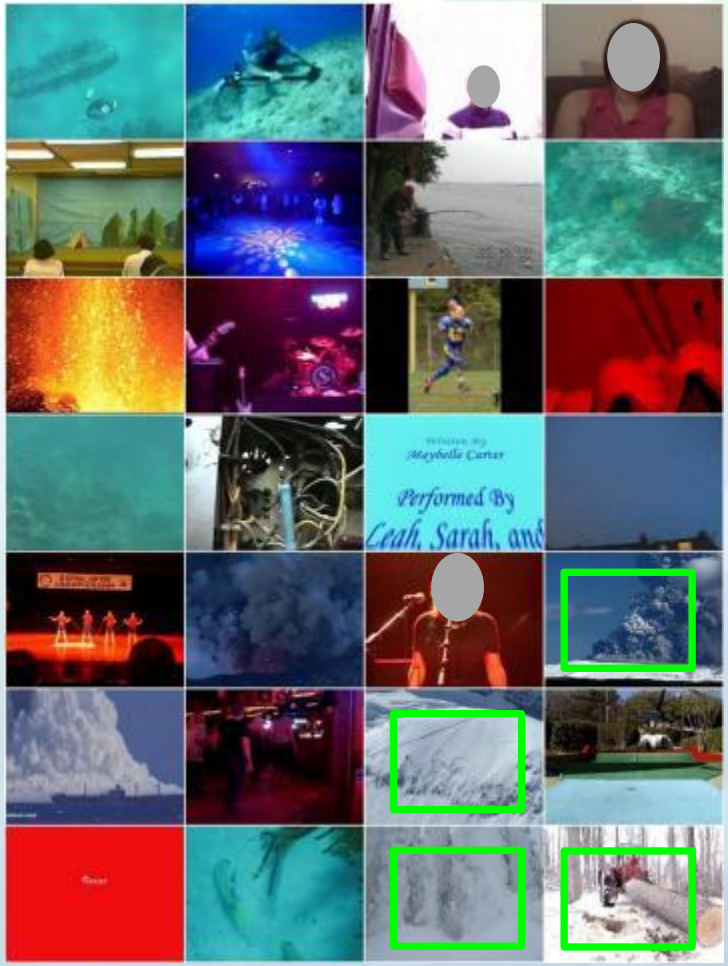


Calibrated vs. Raw Scores: Qualitative Comparison

- Zero shot search (correct retrievals are marked in green)
  - Query: *no horizon* AND *snow*

Calibrated Scores

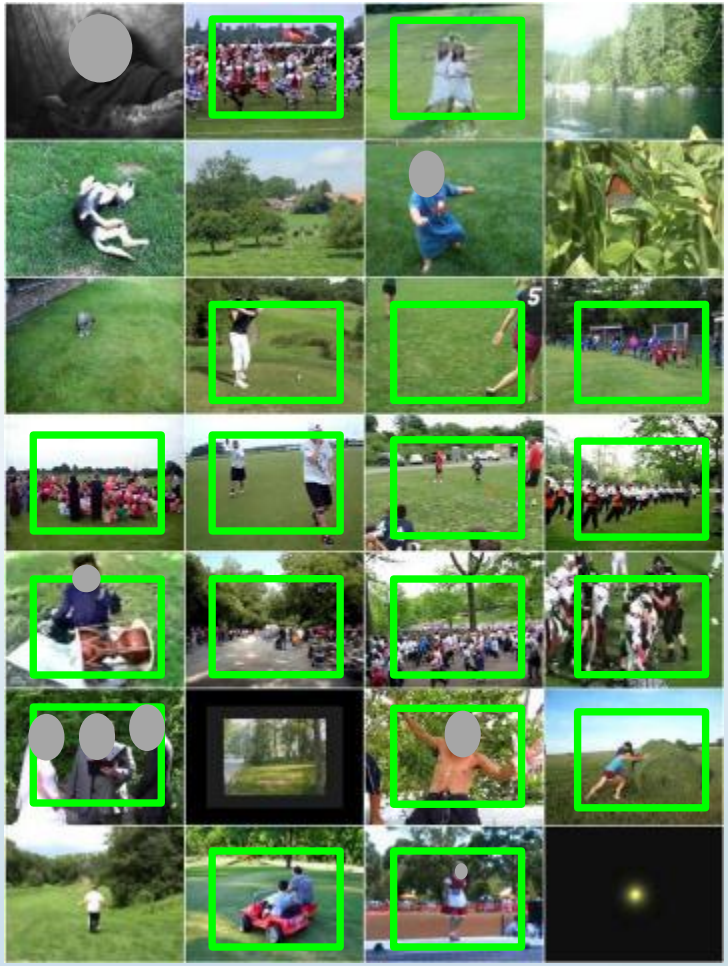
Raw Scores



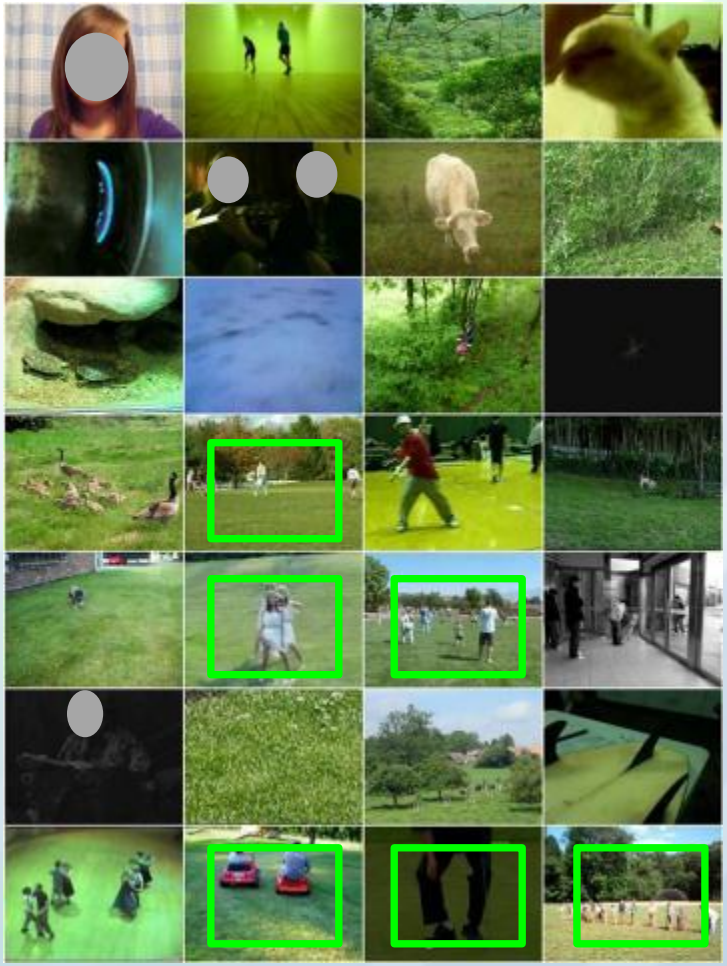
Calibrated vs. Raw Scores: Qualitative Comparison

- Zero shot search (correct retrievals are marked in green)
  - Query: *socializing* AND *vegetation* AND *natural light*

Calibrated Scores



Raw Scores

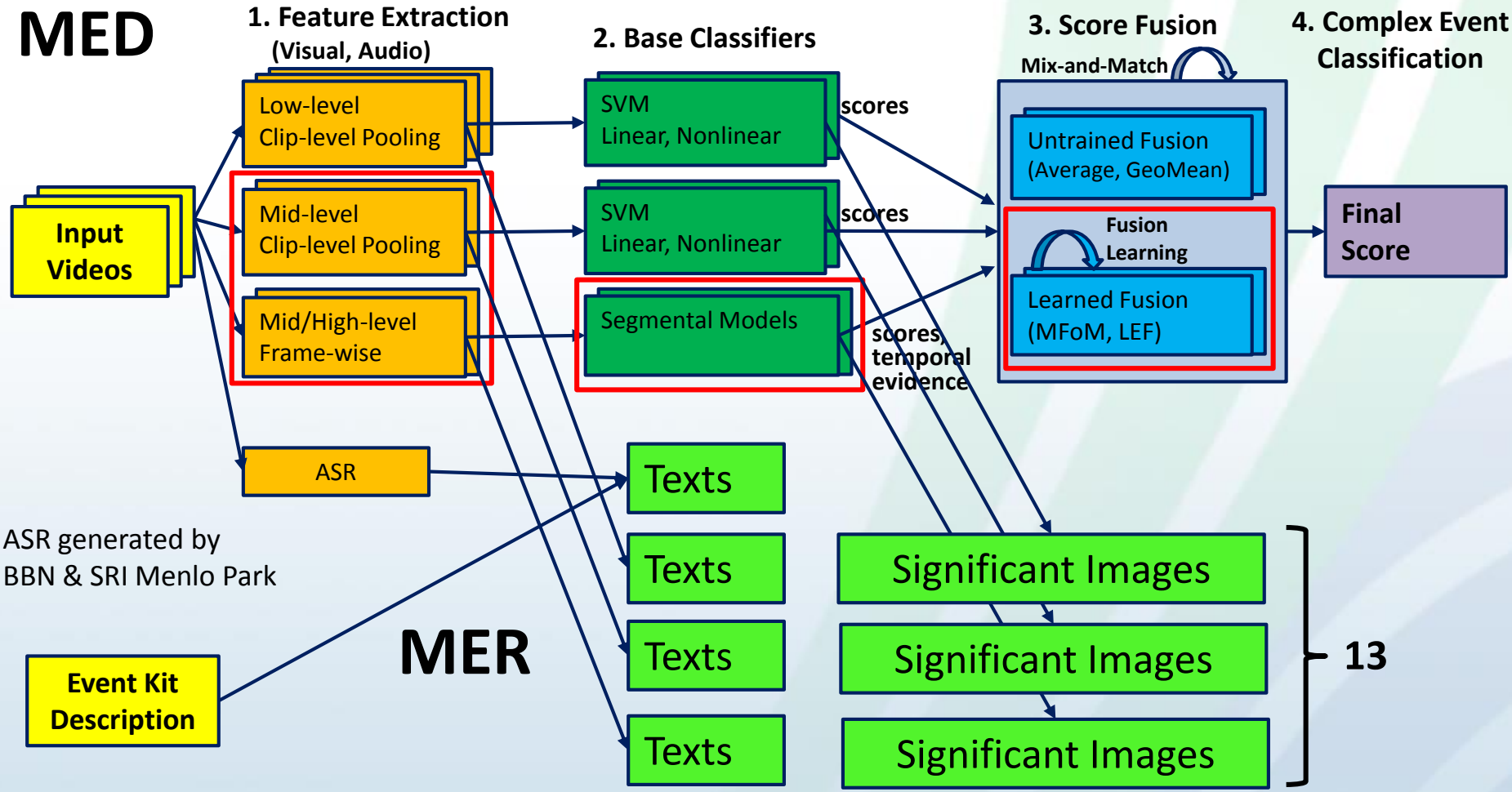


0 Ex

Online Demo

Multimedia Event Recounting

MER Framework



P. Das, C. Xu, R. F. Doell, and J. J. Corso. CVPR '13
"A thousand frames in just a few words:
Lingual description of videos through latent topics and sparse object stitching"

E009 Getting a vehicle unstuck -- HVC701860

- HOG3D_Global_20000_FasterKMeans_xy=9_t=5
  - drive jeep

- HOG3D_Global_4000_xy=9_t=5_Horiztonal_3
  - back camera field push they

E009 Getting a vehicle unstuck -- HVC701860

- op_DenseSIFT
  - back mud pull she woman

- op_geo_color
  - driver jeep mud pickup sand sedan there video

- op_hog2x2
  - it wheel

E009 Getting a vehicle unstuck -- HVC701860

- op_spsift_hesaff
  - camera drive mud vehicl

- op_spsift_mser
  - small then they

- op_ssim

E009 Getting a vehicle unstuck -- HVC701860

- OBDScale
  - coral

- OBD_Avg
  - chair

E009 Getting a vehicle unstuck -- HVC701860

- OBD_Max_Level_3_Horz
  - train, railroad train

- OBD_Max_Temporal_2
  - coral

E022 Cleaning an appliance -- HVC782499

- op_spsift_hesaff
  - cloth demonstr microwav open refriger servic

- op_spsift_mser
  - back brushe clothe end front interview iron move place return rust start tunnel wash women

- op_ssim

E022 Cleaning an appliance -- HVC782499

- HOG3D_Global_20000_FasterKMeans_xy=9_t=5
  - cloth scrub shot vacuum wipe

- HOG3D_Global_4000_xy=9_t=5_Horiztonal_3
  - applianc door kitchen repair shot show water

E022 Cleaning an appliance -- HVC782499

- op_DenseSIFT
  - applianc kitchen microwav open refriger woman

- op_geo_color
  - compani door he open oven shot tray video

- op_hog2x2
  - custom door open process refriger repair stand tray

MER Results

		MER-FullSys		
		Accuracy	ObsTextScore	PRRT
100Ex	AXES	54.18%	1.36	39.17%
	BBNVISER	64.96%	1.78	50.59%
	CERTH-ITI	43.87%	1.06	129.96%
	CMU	55.51%	1.63	52.83%
	Genie	56.44%	0.90	141.23%
	IBM-Columbia	*54.29%	*1.90	*16.57%
	MediaMill			
	NII			
	ORAND			
	PicSOM	64.34%	1.96	36.39%
	SRIAURORA	73.26%	1.58	148.95%
	Sesame	64.10%	2.53	41.83%
	SiegenKobeMuro			
	TNO			
	TokyoTechCanon			
	UMass			
	VIREO	**36.91%	**2.06	**22.93%
	VisQMUL			

Reasonable Accuracy

Image selection might have helped

Many texts are incorrect

Better translation model is needed

Too much information slows down Recounting

Selection or fusion of information needed.

Summary

- TRECVID community is making impressive progress on solving multimedia event detection
  - 100 Ex works well for many event types
  - 10 Ex is the next real challenge, which is gradually being addressed.
  - Search by Semantics (0Ex) is still challenging
- Our attempts to improve 0Ex include:
  - Bottom-Up Concepts
  - Score Calibration
- Our MER incorporated
  - Feature-to-text translation
  - Discriminative Image selection
  - Fusion of results across multiple base classifier is the next step.

Thanks!

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 and by the Defence Advanced Research Projects Agency (DARPA) under contract number HR0011-08-C-0135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, DARPA, or the U.S. Government.